Consilio Institute: Practice Guide

THE GRAND SCAVENGER HUNT: COLLECTION FUNDAMENTALS

Matthew Verga
Director of Education
with additional contributions from Lorraine Moise, Senior Manager, DFES

Consilio / ADVANCED LEARNING INSTITUTE

Consilio / ADVANCED LEARNING INSTITUTE

Consilio Institute: Practice Guide

THE GRAND SCAVENGER HUNT: COLLECTION FUNDAMENTALSS

CONTENTS

THE IMPORTANCE OF COLLECTION

Since electronically-stored information (ESI) has become the norm in discovery, competence with technology has become an essential part of being an effective legal practitioner. With source types multiplying – including challenging sources like mobile devices, social media, and collaboration tools, it is more important than ever for legal practitioners of all types to familiarize themselves with the fundamentals of collection so that they can assist in spotting potential issues and identifying appropriate solutions.

Collection and the Duty of Competence

Understanding the fundamentals of collection is also essential to fulfilling a lawyer's duty of technology competence, which exists in some form in forty states.[1] For example, as articulated in California's Formal Opinion No. 2015-193,[2] there are nine core requirements that lawyers must satisfy to fulfill their duty of technology competence for eDiscovery, two of which explicitly discuss collection: "advise the client on available options for collection and preservation of ESI" and "collect responsive ESI in a manner that preserves the integrity of that ESI." Another four of those nine requirements also necessitate an understanding of collection for their fulfillment ("initially assess e-discovery needs and issues, if any," "implement/cause to implement appropriate ESI preservation procedures," "analyze and understand a client's ESI systems and storage," and "identify custodians of potentially relevant ESI").

Thus, understanding the fundamentals of collection is essential to fulfilling a lawyer's duty of technology competence for eDiscovery in California and, likely, in many other states as well.

ABOUT THIS PRACTICE GUIDE

In this practice guide, we will discuss collection fundamentals, including the broad scope of collection, how computers store ESI, collecting and recovering ESI from computer storage, the intersection of technical and legal realities, self-collection and its risks, in-person and remote collections, and other important collection sources.

THE BROAD SCOPE OF COLLECTION

The practical scope of ESI collection is determined both by the actual requests from other parties and by your own information needs related to the matter. The maximum-possible scope is established by the Federal Rules of Civil Procedure (FRCP) or your state's equivalent ruleset. The FRCP establishes that scope as encompassing:

- Any documents or electronically-stored information
- In your possession, custody, or control
- That are relevant
- That are unique
- That are not unreasonably inaccessible because of undue burden or cost
- That are not disproportionate to the needs of the case

[1]Robert Ambrogi, *Tech Competence*, LAWSITES, https://www.lawsitesblog.com/tech-competence (last visited July 2, 2021).
[2]The State Bar of California Standing Committee On Professional Responsibility and Conduct, *Formal Opinion No. 2015-193* (June 30, 2015), available at https://www.calbar.ca.gov/Portals/0/documents/ethics/Opinions/CAL 2015-193 %5B11-0004%5D (06-30-15) - FINAL.pdf.

The first three criteria set a very broad potential scope for discovery collection. The definition of "documents or electronically stored information" provided by FRCP 34 and its accompanying committee notes[3] is expansive enough to encompass almost any sort of material in any format. "[P]ossession, custody, or control" means that you are responsible, not just for the materials you physically or electronically possess, but for any that you legally control (or, potentially, that you have the practical ability to obtain). "Relevant" is also defined broadly, by Federal Rule of Evidence 401,[4] which states that evidence is relevant if "it has any tendency to make a fact more or less probable than it would be without the evidence" and "the fact is of consequence in determining the action."

The last three criteria set some reasonable, fact-specific limits on that very broad scope. Uniqueness as a limiter comes from the inherently duplicative nature of ESI and from FRCP 26(b)(2)(C)(i)'s admonition that discovery not be "unreasonably cumulative or duplicative." The recognition that some ESI may not need to be produced because it is "not reasonably accessible because of undue burden or cost" comes from FRCP 26(b)(2)(B)[5] (e.g., older data from legacy systems). And, the requirement that all discovery be "proportional to the needs of the case" comes from the 2015 amended[6] definition of the discovery scope itself in FRCP 26(b)(1).[7]

The Technological Scope of Collection

Technologically, this scope means that nothing can be overlooked based purely on its file format or its source type. If it falls within the legal scope described above, you may need to collect it to satisfy a party's request or your own information needs, regardless of whether it comes from:

- **Enterprise systems** (e.g., email, backup, or document management systems) or **departmental systems** (e.g., payroll, research, or compliance systems)
- **Employee computers** (e.g., organization-issued laptops or desktops)
- **Employee storage media** (e.g., thumb drives or external hard drives)
- **Employee mobile devices** (e.g., organization-issued smartphones and tablets or authorized employee-owned devices in BYOD organizations)
- **Cloud-based services** (e.g., storage services, social media services, collaboration tools)
- **Third-party service providers** (e.g., outsourced benefits management)

Collection is not necessarily limited to these common sources either. When the circumstances have warranted it, collection has been necessary from uncommon sources such as vehicle data systems,[8] wearable fitness trackers,[9] and even ephemeral data[10] (*i.e.*, data generated and stored in memory only temporarily as part of a computer system's normal operation). As more and more devices are rendered "smart" and internet-connected, the list of potential sources will continue to grow. For example, photocopiers are almost all networked computers with internal hard drives[11] that store potentially-discoverable copies of the documents they've handled.

[3] Fed. R. Civ. P. 34, available at https://www.law.cornell.edu/rules/frcp/rule_34; Fed. R. Civ. P. 34 advisory committee's note, available at https://www.law.cornell.edu/rules/frcp/rule_34.
[4] Fed. R. Evid. 401, available at https://www.law.cornell.edu/rules/fre/rule_401.
[5] Fed. R. Civ. P. 26(b)(2)(B), available at https://www.law.cornell.edu/rules/frcp/rule_26.
[6] Karen A. Henry and Diana Palacios, The *2015 Amendments to the Federal Rules of Civil Procedure*, AMERICAN BAR ASSOCIATION, https://www.americanbar.org/groups/litigation/committees/minority-trial-lawyer/articles/2016/2015-amendments-to-federal-rules-of-civil-procedure/ (Mar. 1, 2016).
[7] Fed. R. Civ. P. 26(b)(1), available at https://www.law.cornell.edu/rules/frcp/rule_26.
[8] David Horrigan, *e-Discovery Spoliation in Unusual Places: Preserve Your Pickup Truck*, THE RELATIVITY BLOG, https://www.relativity.com/blog/e-discovery-spoliation-in-unusual-places-preserve-your-pickup-truck/ (Mar. 2, 2017).
[9] Samuel Gibbs, *Court sets legal precedent with evidence from Fitbit health tracker*, THE GUARDIAN, https://www.theguardian.com/technology/2014/nov/18/court-accepts-data-fitbit-health-tracker (Nov. 18, 2014).
[10] Kenneth J. Withers, *"Ephemeral Data" and the Duty to Preserve Discoverable Electronically Stored Information*, 37 Univ. of Baltimore L. Rev. 349 (2008), available at https://scholarworks.law.ubalt.edu/ublr/vol37/iss3/4/.
[11] Federal Trade Commission, *Digital Copier Data Security: A Guide for Businesses*, https://www.ftc.gov/tips-advice/business-center/guidance/digital-copier-data-security-guide-businesses (July 2017).

HOW COMPUTERS STORE ESI

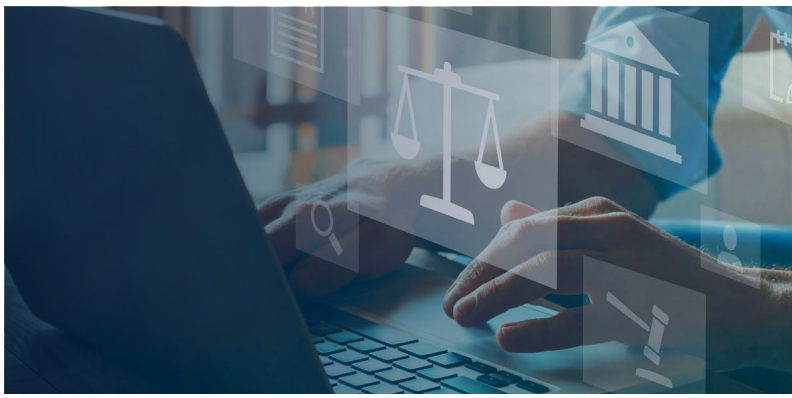
To operate efficiently, computers need to be able to access and work with lots of stored information as quickly as possible:

- Some information is needed to tell a computer's components how to work together
- Some is needed to run the operating system and your applications
- Some is needed to track and respond to your inputs
- Some is needed to retain all of your activity and files

In addition, some of that information needs to be stored reliably even when the computer is off, and some of it is only needed temporarily when the computer is on and performing specific operations. Some of it never changes, and some changes all the time.

As with most things, some memory technologies are fast and expensive, while others are slow and inexpensive. Some of those technologies are volatile, requiring power to maintain storage; others are non-volatile, maintaining storage without power. A mixture of all these memory types is used to satisfy operational requirements while striking a balance between efficiency and affordability:

- Random Access Memory (RAM)
  - Fast, volatile memory that the computer uses for temporary storage of information in active use, including parts of the operating system, parts of applications, and open user files
  - Dynamic RAM (DRAM) is used for the "RAM" component of most computers
- Cache
  - Fast, volatile memory that the central processing unit (CPU) and other computer components use to store information for rapid access to speed up tasks
  - Most computers include two levels of CPU cache, and many now include 3, as well as caches for the graphics processing unit (GPU) and the storage drives
  - Static RAM (SRAM), which is faster but more expensive than DRAM, is often used for these caches
- Storage Drives
  - Slow, non-volatile memory that is used for the bulk of information storage, including the operating system, applications, and all user files and data
    - This is the memory from which collection is most often performed
  - Storage drives can be traditional hard disk drives (HDDs), which work like rewritable record players, or newer solid state drives (SSDs), which cost more but are faster and have none of the moving parts required for HDDs
    - Many computers employ both types of storage drive: a faster SSD for the operating system and key software and a larger HDD for files and media
  - Portions of storage drive memory may also be used as an extension of RAM, known as virtual memory, to further enhance operating efficiency
- Read Only Memory (ROM)
  - Fast, non-volatile memory that contains essential instructions for the operation of the components in the computer



This multi-type, multi-tier approach to memory and storage is also employed in computing devices beyond laptops and desktops. For example, smartphones and tablets employ similar tiered memory systems for the same reasons.

Memory in Motion

As your computer or mobile device operates, there is a constant flow of information being read from and written to storage drives, RAM, and the various caches. At any given moment, multiple copies of a file or portions of a file may exist in multiple locations. These temporary copies are known as ephemeral data, since it typically only exists as long as the computer is on and the operation is active. Collections from individuals' computers and mobile devices are typically only concerned with the static ESI on the storage drive(s), but the ephemeral data generated by enterprise systems has occasionally been implicated in legal matters.[12]

Keeping Track of What's There

Whether a computer or mobile device is using an HDD, an SSD, or both, it is managing a collection of thousands of discrete files that are constantly evolving as files are read, modified, written, and deleted. The computer's file system dictates how this occurs, and although there are a variety of file systems in use in different types of computers and servers, the underlying principles are the same for our purposes.

The immense volume of available storage is divided up into very small physical and logical units. The smallest physical subdivision of a drive is typically referred to as a sector, and some common systems refer to the smallest logical subdivision of the data stored there as a cluster. The specific nomenclature and the specific relationship between physical and logical units depend on the file system in use. Regardless, the computer tracks all of those sectors and clusters in what is, essentially, an enormous spreadsheet that records where each item has been stored and where there is free space to put new things.

Almost all files will be large enough to occupy multiple physical sectors, but those sectors will not necessarily all be physically adjacent. Most of the time, they are spread out across the physical storage, connected only by the entries in the computer's master storage spreadsheet cataloging their relationship. When files are deleted, one of two things happens, depending on the type of storage drive.

In a traditional, platter-based drive, the physical sectors are not wiped clean of their file fragments; rather, the master spreadsheet is just updated to delete the references to that file and to show that those sectors are available once more. In a solid-state drive, the actual data will also be deleted to prolong the life of the drive.

COLLECTING AND RECOVERING ESI FROM COMPUTER STORAGE

As discussed above, we are generally concerned in collection with the primary, non-volatile data storage in a digital device, whether in the form of HDDs, SSDs, or both. On SSDs, what the computer says is there and what's actually there are the same. On HDDs, there is a distinction between what's actually, physically stored on a drive and what the computer is currently tracking in its master storage spreadsheet for that drive, as noted above. This results in two collection options for such drives: physical and logical.

[12] Kenneth J. Withers, *"Ephemeral Data" and the Duty to Preserve Discoverable Electronically Stored Information*, 37 Univ. of Baltimore L. Rev. 349 (2008), available at https://scholarworks.law.ubalt.edu/ublr/vol37/iss3/4/.

Physical collections of HDD storage drives capture an exact copy – or image – of everything on the physical storage, regardless of what the master storage spreadsheet says about where data is and isn't on the drive. This is a bit-by-bit copy, also known as a bitstream copy, which replicates all the physical contents of the storage exactly as they are, essentially creating a virtual duplicate of that physical hardware. The primary benefits of this approach are its completeness and the potential it provides for recovery of deleted files.

Logical collections of HDD storage drives work within the file system's management of the storage drive rather than replicating the whole piece of hardware. Logical images exactly replicate everything tracked in the computer's master storage spreadsheet or some defined subset of it (e.g., everything in particular directories or folders). The primary benefit of this approach is the potential to target more narrowly and collect less extraneous material, such as non-reviewable system files.

Recovery of Deleted Files

On traditional platter-based hard drives, there are two potential sources of information that can be captured in physical images that are not captured in logical ones: slack space and unallocated space. As we noted above, files in computer storage take up multiple sectors or clusters on a drive. Sectors or clusters not currently in use for active storage are referred to as unallocated space. Some sectors or clusters that are in active use may only be partially full. The remaining, unused portion of the sector or cluster is referred to as slack space.

On platter-based hard drives, computer deletion only deletes the records of what's in sectors and clusters, rather than actually erasing them, so both unallocated space and slack space may contain fragments of deleted files that had been stored there. A forensic examiner working with a full, physical image may be able to use specialized software tools to recover files or file fragments from unallocated or slack space and render them usable for investigation or litigation. While this is not a typical step in routine eDiscovery work, and the files must be reviewed and interpreted by a forensic expert rather than an attorney, it may be worth the effort in some cases.

In recent years, however, a transition has occurred from platter-based hard drives being the norm to solid-state hard drives being the norm. Most new computers and devices are now built using SSDs instead of traditional platter-based hard drives. Unlike the older drives, SSDs are designed not to retain anything in unused storage space to prolong the life of the drive. So, for any newer digital devices, recovery of deleted files may not be possible.

Preventing Alteration During Collection

During the collection of ESI from computer or device storage drives, it is important to avoid altering the source in any way by the act of collection. As we noted above, computers are designed for efficiency rather than data preservation, and when in operation, they have a constant flow of information being read from and written to various memory components. To avoid doing any new writing to a drive during the act of reading from a drive, forensic examiners use tools called write blockers. Write blockers are specialized hardware or software tools that block any write commands from being passed to a drive while it is being accessed for collection, ensuring the original source is unaltered by the collection activity.

Verifying the Accuracy of Collection

Just as important as avoiding alteration to the source is verifying that the copies you've made are accurate ones. When copying large volumes of ESI (300,000 to 500,000 files per source drive is common), there is some potential for errors to occur during the copying of some of those files.

Hashing is used to validate that all files have been copied accurately.

Hashing is a technique by which sufficiently unique "fingerprints" can be generated for files. Hash functions are mathematical processes that take irregular-length inputs (e.g., the data in a particular file), and use them to generate fixed-length outputs (e.g., a string of 32 numbers and letters). In collection, hashing is typically accomplished using a cryptographic hash function (e.g., MD5 or SHA-1), which is well-suited to matching unique inputs to particular outputs.

To verify a collection's accuracy, one set of fingerprints is generated from the source files, and that set is then compared to a second set generated from the copied files. Fingerprint matches confirm an accurate copy, and fingerprint mismatches identify copying errors.

THE INTERSECTION OF TECHNICAL AND LEGAL REALITIES

The ultimate goal of evidence collection is the eventual use of some of that evidence in court, whether by you or another party. The admissibility of a particular piece of evidence at trial turns on a variety of factors, including its relevance, its potential for prejudice, its status as hearsay, etc. The most foundational requirement offered evidence must satisfy is that it must be authentic, *i.e.* it must actually be whatever it purports to be. This is essential for the obvious reason that fake or falsified or altered materials cannot carry any weight as evidence. Fake evidence makes no fact more or less true and is, therefore, irrelevant to the proceedings.[13]

The process for establishing evidentiary authenticity is laid out in Federal Rule of Evidence 901.[14] To establish authenticity, "the proponent must produce evidence sufficient to support a finding that the item is what the proponent claims it is." Satisfying this requirement for ESI means being able to demonstrate that an offered file comes from where you say it does and has not been altered from the original, *i.e.* that you've maintained forensic soundness and chain of custody.

Forensic Soundness

Forensic soundness is a widely used phrase in the discussion of forensic collection and investigation processes that lacks a precise legal or technical definition. It is used generally to describe tools and processes that can be relied upon to capture evidence in a way that does not alter or corrupt that evidence, and which conforms to accepted industry best practices. For working with ESI, the National Institute of Standards and Technology actually tests the operation of available forensic tools (like the write blocking and disk imaging tools mentioned above) and provides public reports on their soundness.[15]

In the context of eDiscovery, ensuring forensic soundness generally means capturing exact copies of relevant files, with any relevant metadata intact, and then working with copies of those copies, to ensure preservation of an unaltered original set. The precise technical steps required to achieve that goal will vary by ESI source and collection tools employed, and the currently-accepted industry best practices for various source types continue to evolve as the technology does, both in practice and in court. For this reason, engaging the services of a qualified forensic expert – or at least consulting with one prior to collection – is recommended to ensure currently-accepted tools and processes are employed.

[13] Fed. R. Evid. 401, available at https://www.law.cornell.edu/rules/fre/rule_401.
[14] Fed. R. Evid. 901, available at https://www.law.cornell.edu/rules/fre/rule_901.
[15] *Computer Forensics Tool Testing Program (CFTT)*, NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY, https://www.nist.gov/itl/ssd/software-quality-group/computer-forensics-tool-testing-program-cftt (Nov. 15, 2019).

The Importance of Maintaining Metadata

Metadata, broadly speaking, is data about data. In the context of ESI, every file on a computer or mobile device contains not only the primary content you would see if you opened it (e.g., the body of an email) but also a diverse array of information about the file itself. Common examples include the time and date sent for an email, or the author and last modification date for an Office document. Some kinds of metadata are visible to users of those applications, some can be viewed by viewing a file's properties in Windows or MacOS, and others are not typically visible to users. All of this additional information is the file's metadata, and it is an important part of collection and discovery.

The specific metadata fields available will vary with the specific file format. For example, music files typically include artist and track information in their metadata. Photo files may record where they were taken and by what device. Email files will document their attachments. Application metadata ranges from the very widely-used (e.g., date and time created) to the very application-specific (e.g., tracked changes in a document or hidden content in a spreadsheet). Additional metadata about files may also come from the system on which they exist (e.g., file path).

In terms of evidentiary value, we are most often concerned with metadata revealing when things were done (e.g., when something was created, modified, sent, or received), but there may be relevant evidence in other types of metadata, and there is enormous process value regardless. Metadata values are the basis of many filtering, sorting, and searching options within document review tools. For example, metadata is what enables you to keep emails and attachments in family groups, to filter for emails to or from a particular address, or to search for keywords within email subject lines. The more data about your data you have, the more creative and efficient you can be in your exploration of that data during early case assessment and review.

Because of both its potential evidentiary value and its enormous process utility, metadata has become an expected and sometimes required[16] component of many ESI productions, such as DOJ productions.[17] Unfortunately, metadata is also easily altered if files are not collected and handled correctly. For example, accessing and copying original files without safeguards like those we discussed above can alter metadata, as can forwarding relevant emails[18] instead of collecting them directly. Such alterations to metadata would destroy forensic soundness, reduce utility, and potentially, impair the admissibility of the evidence and the defensibility of your discovery.

Chain of Custody

Chain of custody refers to documentation of the path a piece of evidence has traveled from its point of origin to its eventual submission in court. It typically documents places, times, and people involved in the handling of the evidence, as well as any relevant processes employed. Its purpose is to demonstrate that a piece of evidence submitted in court is what you claim it is, from where you claim it's from, and unaltered, as required by Federal Rule of Evidence 901.[19]

[16] *Singh v. Hancock Natural Resources Group, Inc.*, 2016 WL 7474886 (E.D. Cal. Dec. 29, 206), available at https://scholar.google.com/scholar_case?case=2059425785764292803.
[17] Antitrust Division, *Electronic Production Letter (Attachment 1)*, U.S. DEPT. OF JUSTICE, https://www.justice.gov/atr/electronic-production-letter-attachment-1 (June 2015).
[18] *Singh*, supra note 16.
[19] Fed. R. Evid. 901, available at https://www.law.cornell.edu/rules/fre/rule_901.

Although the concept originates with physical evidence, it is equally applicable to ESI collection and handling. Thus, carefully documenting your collection efforts and subsequent ESI handling is another important part of ensuring the reliability and later admissibility of the ESI you collect. In addition to your chain of custody documentation, an individual responsible for the collection and data handling may need to submit an affidavit (or provide live testimony) describing the steps taken, the tools used, and how forensic soundness and chain of custody were both maintained from the point of collection to the point of submission as evidence.

SELF-COLLECTION AND ITS RISKS

Custodian self-collection refers to a collection approach in which the custodians themselves undertake the identification and collection of relevant documents from their own materials. For example, they might review their physical records and turn over any relevant paper files to a designated recipient in the in-house counsel's office, or they might review their stored electronic files and place copies of relevant materials in a designated storage area on the organization's network or to a designated folder in Outlook.

Custodian self-collection of ESI carries four categories of risk that can each lead to spoliation sanctions, authentication and admissibility issues, and other negative consequences, which is what makes custodian self-collection approaches unsuitable for almost all matters:

- **Generic Inaction**: The first category of risk you run when leaving collection to the custodians is that they simply may not do it. Employees are busy doing their normal job duties, and most do not understand the importance of preservation and collection the way lawyers do. It is not uncommon to have to chase employees down just to get them to acknowledge receiving a legal hold. Asking them to execute a complex, time-consuming collection process is likely to go right to the bottom of their to-do list. And, even if you eventually get everyone to act on your instructions, the delays before action can lead to the loss or alteration of relevant materials through normal work activities, automated janitorial processes, or system or device failures.
- **Legal Misunderstanding**: The second category of risk you run when leaving collection to the custodians is that they will misunderstand or misapply the legal and factual scope information you give them in your instructions. The scope of preservation and collection is defined through the interaction of a nuanced legal standard, the pleadings and discovery requests of the parties, and the facts known at the time. The scope of relevance (and, thus, of collection) frequently evolves over the course of discovery as legal disputes are refined and more factual knowledge is gained. Expecting non-lawyer employees to clearly understand nuance with which lawyers frequently struggle is a recipe for disappointment, and expecting that nuance to be consistently applied from employee to employee is even more so. And, when employees misunderstand or misapply the scope you've tried to set, relevant materials can end up omitted or lost altogether.
- **Technical Ineffectiveness**: The third category of risk you run when leaving collection to the custodians is that, even if they perform the requested collection and apply the scope guidance as you intended, they may still execute the process in a technically ineffective manner resulting in materials being missed, lost, or altered. For example, custodians asked to run searches to locate their relevant materials may design those searches ineffectively or execute provided ones incorrectly, causing relevant materials to be missed entirely. Minor changes to search syntax or search settings can make major differences in the results returned, and syntax and settings vary from system to system.

Moreover, ESI materials and their metadata are easily altered by almost any interaction with a file. Custodians working without write blockers or other forensic tools cannot maintain forensic soundness or perform hash validation. Some metadata will be altered, which may affect the ESI's evidentiary value, its authentication, or its admissibility.

- **Intentional Misconduct**: The final category of risk you run when leaving collection to the custodians is that they will engage in intentional omission, alteration, or destruction of materials to conceal their own actions. There are many situations in which your custodians' interests may run counter to your organization's. For example, they may be responsible for some part of the events giving rise to the matter and may fear getting in trouble themselves, they may be engaged in some unrelated misconduct they are afraid may be exposed, or they may think they're protecting a colleague or the organization. Whatever the reason, when custodians are trusted to self-collect ESI they have the opportunity to commit sins of omission or spoliation. And, even if they do not take that opportunity, another party may challenge the reliability of collection performed by a custodian with an individual interest in the matter or the materials.

One common variation on custodian self-collection is organization self-collection, which refers to an approach in which an organization leverages its information technology personnel to perform collection of ESI. For example, the administrator of the organization's email system may perform searches and exports from that system, or IT personnel might be directed to image specific employees' work computers. The materials collected by IT are then typically turned over to a law firm or a discovery services provider for subsequent processing, hosting, review, and production.

While organization self-collection is usually a less risky approach than custodian self-collection, it is still a risky approach for the same reasons listed above. Additionally, it has the potential to be both expensive and disruptive to the normal operations handled by the repurposed personnel.

The Courts on Self-Collection

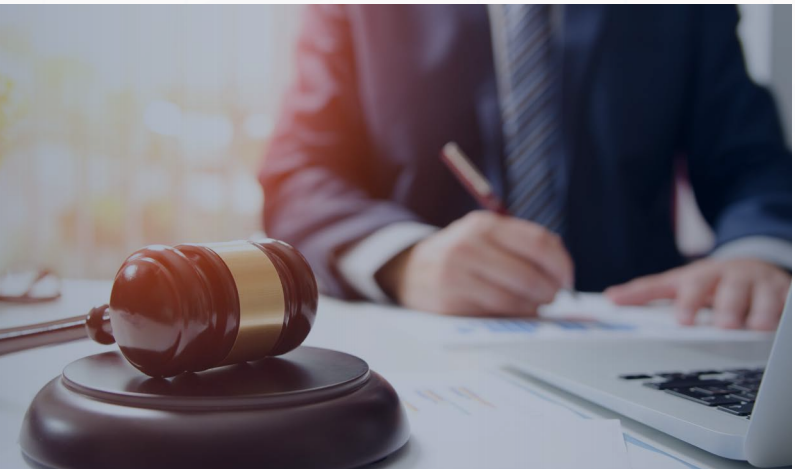
The risks and consequences of employing self-collection approaches are not merely hypothetical. For many years, courts have highlighted those risks, have taken parties and their lawyers to task for their reliance on self-collection in the face of those risks, and have applied significant monetary and evidentiary sanctions for failures caused by taking those risks:

- *Leidig v. BuzzFeed, Inc.*, 2017 WL 6512353 (S.D.N.Y. Dec. 19, 2017)[20]
  - In this case, an "amateurish collection of documents [led] to the destruction of perhaps critical metadata." The metadata was "irreversibly destroyed" when the plaintiff himself "transferred the files to a new device." As a result of this spoliation caused by self-collection, the plaintiff was precluded from using the dates of the affected documents as evidence.
- *National Day Laborer Organizing Network, et al., v. U.S. Immigration and Customs Enforcement Agency, et al.*, 877 F.Supp.2d 87 (S.D.N.Y. Jul. 13, 2012)[21]
  - In this case, defendant government agencies had collection searches performed by individual custodians with no meaningful direction or oversight of their searching. Moreover, most of the custodians' search efforts were undocumented, making post hoc evaluation of their adequacy impossible. As a result of this custodian self-collection process, the defendants were ultimately directed to undertake

[20] Available at https://casetext.com/case/leidig-v-buzzfeed-inc.
[21] Available at https://casetext.com/case/natl-day-laborer-organizing-network-v-us-immigration-customs-enforcement-agency.

significant additional discovery work to ensure acceptable quality and completeness would be achieved.

- *Peter Kiewit Sons', Inc. v. Wall Street Equity Group, Inc., et al.*, 2012 WL 1852048 (D. Neb. May 18, 2012)[22]
  - In this case, the defendants were sanctioned for a host of discovery failures, several of which were the result of relying on employees for management of ESI sources and for conducting the searches for relevant ESI, leading the court to conclude that the "Defendants' search of their electronic files to provide discovery responses was woefully inadequate." Ultimately, the magistrate judge concluded that the search process employed by the defendants was not "a good faith search for the electronically stored information." As sanctions, the court ordered the defendants to pay attorney fees and expenses for the motion practice and additional discovery and imposed a permissive adverse inference jury instruction.
- SunTrust Mortgage, Inc. v. AIG United Guaranty Corp., et al., 2011 WL 1225989 (E.D. Va. Mar. 29, 2011)[23]
  - In this case, the plaintiff relied upon an employee central to the underlying dispute to perform identification and collection of relevant materials, and that employee took the opportunity to alter several relevant documents to make them support her version of events. In-house and outside counsel later relied upon an email they did not know had been altered to support their amended complaint. Ultimately, additional altered emails were discovered, and the court found that the employee had perpetrated a fraud on the court for which the plaintiff was responsible. As sanction for the fraud and abuse, the plaintiff was ordered to pay defendant's "very significant additional legal fees and expenses" that were incurred to "preserve[] the integrity of the judicial record."
- *Green v. Blitz U.S.A., Inc.*, 2011 WL 806011 (E.D. Tex. Mar. 1, 2011),[24] *vacated after settlement*, 2014 WL 2591344 (E.D. Tex. June 1, 2014).
  - In this case, the defendant's discovery workflow involved a particular employee within the company meeting with counsel to find out what materials might be relevant and then talking to individuals or departments he thought might have such materials and asking them to provide him with those materials. The employee "did not institute a litigation-hold of documents, do any electronic word searches for emails, or talk with the IT department regarding how to search for electronic documents." As a result, numerous relevant documents were missed and never produced – some of which could have been found with "shocking . . . ease" if appropriate IT or collection experts had been involved in the process. On the basis of their significant, "willful" discovery failures, the court ordered the defendant to pay a $250,000 civil contempt sanction and to furnish a copy of the order "to every Plaintiff in every lawsuit it has had proceeding against it" for the past two years and the next five years.



[22] Available at https://www.govinfo.gov/content/pkg/USCOURTS-ned-8_10-cv-00365/pdf/USCOURTS-ned-8_10-cv-00365-14.pdf.
[23] Available at http://amlawdaily.typepad.com/04142011suntrust.pdf.[24] Available at http://amlawdaily.typepad.com/04142011suntrust.pdf.
[24] Available at https://casetext.com/case/green-v-blitz-usa.

IN-PERSON AND REMOTE COLLECTIONS

Traditionally, the most common collection approach has been in-person collection, in which the person executing the collection is in physical possession of the devices to be collected. This may be achieved by sending the devices to the person executing the collection, but it is more often achieved by having the person executing the collection travel to where the custodians and their devices are.

In-person collection has many benefits. Most importantly, it ensures proper collection from the original source overseen by a professional rather than by the custodian. It can also give the person executing the collection an opportunity to interact with the custodians to gather useful information about the sources and what they contain, combining in-person custodian interviews with collection itself. And, when multiple custodians are in the same office location, it can be efficient, even when travel to that location is required.

In-person collection is not ideal for all situations, however. When custodians are distributed across multiple office locations – or when many employees work remotely from home – travel to all of those locations can quickly become too costly and time-consuming to make sense. Having one or more collection professionals on-site can also be disruptive to normal operations and may not be as subtle an approach as you require in certain investigative contexts.

The Rise of Remote Collection

As geographically-distributed (and remote) employees have become more common, remote collection has grown in popularity as an approach. Remote collection comes in four primary subtypes:

- **Self-Executing Devices with Instructions** – In the first subtype, a collection device is prepared in advance and shipped to the custodian who follows provided instructions to connect the device to their computer and initiate the pre-defined collection process that automatically copies the files it was set up to copy (e.g., specified file types, specified directories, etc.). When the collection is complete, the custodian returns the device.
- **Preconfigured Drives Plus Remote Access** – In the second subtype, which is the most widely used, a preconfigured drive is shipped to the custodian who connects it to the computer and then grants remote access to a remote collection professional to execute and oversee the actual collection to the drive. When the collection is complete, the custodian returns the drive. (It is also possible for limited collection over the internet to be done by remote access, but typically the sizes involved make shipping drives a better choice for this approach.)
- **Preconfigured Laptops Plus Remote Access** – The third subtype is a solution developed for remote collection of smartphone data. In this approach a laptop with the necessary collection software is shipped to the custodian, who connects the laptop to the internet and their smartphone to the laptop. A collection professional can then connect to the preconfigured laptop remotely to perform the required collection. When the collection is complete, the custodian returns the laptop.
- **Enterprise Applications** – In the fourth subtype, an enterprise application is installed on the organization's network environment that facilitates manual or automated collections from devices connected to the network. Collections may be executed with or without the custodians' knowledge and may be administered by IT personnel or by third-party collection professionals. Collections are copied over the network and later may be transferred to drives for shipping to a third-party eDiscovery services provider for processing and review. Due to the cost of such applications, they are most often used by large organizations.

OTHER IMPORTANT COLLECTION SOURCES

Thus far, we have spoken primarily about the collection of ESI materials from the computers of individual custodians, but most cases involve collection from a range of other sources as well. The fundamentals of computer memory operation and successful acquisition from that memory are the same regardless, however you still need to be aware of the other source types you may need to consider and the complications that they entail.

The other major categories of sources are:

- Enterprise systems
- Mobile devices and apps
- Social media platforms
- Collaboration tools

Collection from Enterprise Systems

Enterprise systems refers to the software and hardware systems maintained by your organization or its departments, including email systems, internal instant messaging systems, document management systems, CRM or ERP systems, internal collaboration tools, backup systems, and more. Depending on the nature of the matter, it might also include voicemail systems, security and video systems, or even networked photocopiers or other office machines.

How collection from such systems is performed can vary widely depending on the system. Some systems store their data in ways that can be directly collected like the materials on a custodian's computer, while others require you to use the system's built-in search and export tools. Those tools may have material limitations that affect what results a search can return or what an export can contain. Working closely with the responsible IT personnel to ensure those limitations are understood and accounted for is critical when collecting data from enterprise systems.

Collection from Mobile Devices and Apps

Mobile devices – smartphones in particular – have become ubiquitous for both personal and business life. Like all consumer technology, there are a plethora of models and types available, and new ones are released by each maker each year. And, because many organizations have adopted bring-your-own-device policies (BYOD), organizations may have a much wider variety of smartphones as potential sources than computers (which still tend to be organization-selected and issued).

Smartphones are more difficult, more costly, and more time-consuming to collect and process than computers. The difficulty, cost, and time can vary from model to model, from maker to maker, and from operating system to operating system. Collection directly from smartphones requires specialized tools like those used to collect from a custodian's computer. Collections instead from cloud-based backups of the smartphone in question are sometimes also an option.

Different models run different types of operating systems, and the operating systems differ in functionality and are updated regularly. Updates can affect the way in which applications store their data or how they are backed up. In other words, data that can be forensically extracted today, may not be able to be extracted tomorrow, or vice versa.

At a high level, applications that come pre-installed on a mobile device when you take it out of the box and power it on, such as Contacts, SMS, MMS, Calendar, Photos, and Video, will typically be extracted from the handset during a standard imaging process using forensic tools. These applications are known as "stock" applications.

Third-party applications on mobile devices, which are applications that the user downloads onto the handset from digital storefronts like the Apple App Store or the Google Play Store, may or may not be

extracted from the handset during a standard device collection process. This may be because of end-to-end encryption or other security measures implemented by the app's developers.

This varies not just from app-to-app but even across devices and operating systems. For example, WhatsApp data is stored in an encrypted format on recent Android devices and cannot be extracted as part of a standard mobile phone imaging. This is not the case with iPhone, where WhatsApp data could be captured in a readable format.

Some third-party applications store data within the cloud as opposed to on the user's device. Data from these applications cannot be extracted from a user's device during a standard collection. Applications that store data within the cloud may require separate standalone collections directly from the cloud services.

Additionally, it is important not to overlook less common mobile devices that may, at times, be relevant, such as vehicle GPS or data systems,[25] wearable devices like fitness trackers,[26] etc.

Collection from Social Media Platforms

For better or worse, social media is an influential, indispensable part of modern life. As it's permeated its way ever deeper into our professional and personal lives, its impact upon discovery has grown in parallel. In April 2019, the International Legal Technology Association published the results of its 2018 Litigation and Practice Support Survey,[27] revealing that 90% of responding professionals (overwhelmingly from law firms) had handled at least one case involving the collection and processing of social media data in the prior year, a 7% increase over the prior year.[28] Moreover, 19% reported handling more than 20 such cases, a 46% increase over the prior year.

Social media sources can pose technical challenges because they typically incorporate multiple forms and formats of media and communication together, creating a complex source of diverse ESI. They commonly allow sharing of photos and videos, status updates, public posts, private messages, live chats, video streams, and more. In addition to the material posted and uploaded by users, social media services also record extensive information[29] about each user's activities on the service, such as what content they've liked or shared, logs of when and how they've accessed the service, and sometimes more.

All this material accumulates rapidly into large volumes because social media users access these services frequently and share hundreds of millions of new posts, messages, photos, and videos every day. Each individual social media account for each user can easily contain hundreds or thousands of pages of materials in a mishmash of formats. Facebook, for example, published a paper in 2021 on its transition to a new file system for its data centers[30] in which each cluster "scales to exabytes," up from "tens of petabytes" in their previous system.

There are three main options for the acquisition of social media materials for use in litigation:

- *Printing out the material or capturing a screen image of it* – this is fast and inexpensive, but it does not capture any native files or metadata. It may also create authentication and admission problems down the road.
- *Using the self-service export tools provided by the social media platform* – this, too, is fast and inexpensive, but it also may not provide native files or metadata. It often comes in a format that requires conversion using forensic tools, and not all parts of the content may be exported in a way that facilitates that conversion.

[25] David Horrigan, *e-Discovery Spoliation in Unusual Places: Preserve Your Pickup Truck*, RELATIVITY BLOG, https://www.relativity.com/blog/e-discovery-spoliation-in-unusual-places-preserve-your-pickup-truck/ (Mar. 2, 2017).
[26] Katherine E. Vinez, *The Admissibility of Data Collected from Wearable Devices*, 4 Stetson J. Advoc. & L. 1 (2017), available at https://www2.stetson.edu/advocacy-journal/wp-content/uploads/2017/06/Vinez_-_Wearables.pdf.
[27] Cindy MacBean, *2018 Litigation and Practice Support Survey Results*, ILTA (Apr. 2019), available at http://epubs.iltanet.org/i/1108621-lps19/36?_ga=2.231156186.434461956.1629978821-1135214194.1629978821.
[28] *ILTA's 2017 Litigation and Practice Support Technology Survey Results*, ILTA (Apr. 2018), available at http://epubs.iltanet.org/i/973671-lps18/55?_ga=2.39038435.1141759458.1531162513-441756871.1531162513.
[29] *What categories of my Facebook data are available to me?*, FACEBOOK HELP CENTER, https://www.facebook.com/help/405183566203254?helpref=faq_content (2021).
[30] *Consolidating Facebook storage infrastructure with Tectonic file system*, FACEBOOK ENGINEERING, https://engineering.fb.com/2021/06/21/data-infrastructure/tectonic-file-system/ (June 21, 2021).

- *Using specialized forensic collection software* – this carries additional costs, but it can be essential for cases involving large quantities of social media materials, questions best resolved through the materials' metadata, or the potential for disputes over the authenticity and admissibility of the social media materials themselves. Escalating security and privacy measures, however, have begun to reduce how much these tools can do beyond the standard export function.

Collection from Collaboration Tools and Microsoft 365

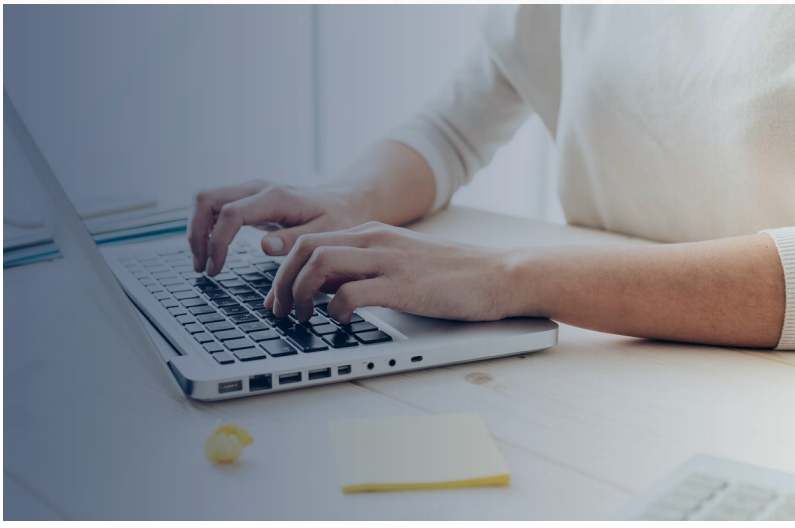
Collection from collaboration tools like Slack and Teams requires navigating a collection of diverse sources, containing diverse content, and potentially, stored in diverse locations. Relevant communications may exist in public channels, private channels, direct messages, or group messages. It is not uncommon for an organization to have channels numbering in the thousands and messages numbering the millions. Moreover, each message may contain reactions, animations, links to videos, embedded content from third-party sources, and more.

Another challenge arises from the variety of licenses available. The type of license under which an organization uses Slack will dictate what options are available for preservation and export of relevant materials. For example, a free license for Slack caps how many messages can be preserved and exported, while paid licenses do not. Paid licenses also allow for more granular preservation options. In Teams as well, the Microsoft 365 license under which an organization uses Teams will determine what preservation and export tools are available.

An additional challenge arises from the diversity of places where relevant data may reside, which can complicate preservation and export. For example, different types of Teams data are stored in different places within the Microsoft 365 environment. Individual Teams content is stored in a user's mailbox, non-private channels content is stored in the group mailbox used for the team, and other types of content are stored in various SharePoint and OneDrive locations. In Slack or Teams, embedded content may be stored in third-party applications (e.g., Dropbox, YouTube) and just displayed dynamically based on the link that's actually in the message.

Because of these challenges and variations, as well as the additional challenges that arise during processing (e.g., expansion, format conversion, unitization), successful collection from these kinds of sources typically requires the assistance of an experienced collection expert, and it may require custom solutions.

Preservation in and export from Microsoft 365 presents the same challenges discussed above for Teams. It encompasses a wide range of sources and date types. It can contain enormous numbers of files and enormous volumes of data. Preservation and export options are dictated by license level, and they are complicated by the diverse array of places different types of user data is stored – both inside the Microsoft 365 environment and in third-party applications. Successful collection from these kinds of sources typically requires the assistance of an experienced collection expert (as well as the cooperation of the account holder, for individual accounts), and it may require custom solutions.



KEY TAKEAWAYS

There are eight key takeaways from this practice guide to remember:

1. Understanding the fundamentals of collection is necessary to successfully navigate discovery and to fulfill lawyers' duty of technology competence.

2. The potential scope of collection is very broad, both legally and technically, and it continually evolves as new devices and services become available and as people's patterns of behavior change to incorporate those things into their work and their lives.

3. The complex nature of computer operations and ESI storage leads to the creation of unneeded duplicates and ephemeral files, as well as the potential for easy alteration and accidental loss (though deleted ESI may sometimes be recoverable).

4. Ensuring that forensic soundness (*i.e.*, no alternation) and chain of custody (*i.e.*, documented path from original source to introduction in court) are maintained during collection is essential to avoiding potential issues with authentication and admissibility.

5. Metadata has both evidentiary and process value, and it is an expected – and sometimes required – component of collection (and later production).

6. Collecting ESI, without alteration and without loss of metadata, requires special tools (e.g., write blockers) and processes and, typically, outside experts in forensic collection.

7. Available collection approaches include: custodian and organization self-collection (high risk), in-person collection (sometimes high cost), and several varieties of remote collection (currently the most popular).

8. When planning collection, don't forget source types beyond individual custodians' computers, including: enterprise systems, mobile devices and apps, social media platforms, collaboration tools, and more (e.g., vehicle systems, wearables, etc.).

ABOUT THE AUTHOR

Matthew Verga is an attorney, consultant, and eDiscovery expert proficient at leveraging his legal experience, his technical knowledge, and his communication skills to make complex eDiscovery topics accessible to diverse audiences. A sixteen-year industry veteran, Matthew has worked across every phase of the EDRM and at every level, from the project trenches to enterprise program design. As Director of Education for Consilio, he leverages this background to produce engaging educational content to empower practitioners at all levels with knowledge they can use to improve their projects, their careers, and their organizations.

Matthew Verga

Director of Education

m +1.704.582.2192

e matthew.verga@consilio.com

consilio.com